14452-75-6; GaH, 13572-92-4; GeH, 13572-99-1; AsH, 12628-08-9; BrH, 10035-10-6; Be, 7440-41-7; B, 7440-42-8; C, 7440-44-0; N, 17778-88-0; O_2, 17778-80-2; Na, 7440-23-5; Mg, 7439-95-4; Al, 7429-90-5; Si, 7440-21-3; P, 7723-14-0; S, 7704-34-9; Cl, 22537-15-1; K, 7440-09-7; Ca, 7440-70-2; Ga, 7440-55-3; Ge, 7440-56-4; As, 7440-38-2; Se, 7782-49-2; Br_2, 10097-32-2; HOCl, 7790-92-3; SeH, 13940-22-2; BeH_2, 7787-52-2; BH_3, 13283-31-3; CH_4, 74-82-8; CH_3CH_3, 74-84-0;

$CH_2=CH_2$, 74-85-1; $CH\equiv CH$, 74-86-2; HCN, 74-90-8; CH_2O, 50-00-0; CHO, 2597-44-6; CH_3F, 593-53-3; CO_2H, 64-18-6; CH_3OH, 67-56-1; NH_3, 7664-41-7; HNC, 6914-07-4; HNO, 14332-28-6; HNO_2, 7782-77-6; HNCO, 75-13-8; H_2O, 7732-18-5; HOCN, 420-05-3; HOCHO, 64-18-6; MgH_2, 7693-27-8; AlH_3, 7784-21-6; SiH_4, 7803-62-5; PH_3, 7803-51-2; SH_2, 7783-06-4; GeH_4, 7782-65-2; AsH_3, 7784-42-1; SeH_2, 7783-07-5.

Molecular Modeling of the Physical Properties of the Alkanes

Diane E. Needham, I-Chien Wei, and Paul G. Seybold*

Contribution from the Departments of Chemistry and Biochemistry, Wright State University, Dayton, Ohio 45435. Received August 27, 1987

Abstract: Eight physical properties (boiling points, molar volumes, molar refractions, heats of vaporization, surface tensions, melting points, critical temperatures, and critical pressures) of 74 normal and branched alkanes were examined by molecular modeling techniques. Structural parameters employed include Wiener indices, connectivity indices, ad hoc descriptors, information indices, and molecular volumes and surface areas. Most of the properties were well modeled ($r^2 > 0.97$) by the Wiener indices, connectivity indices, and ad hoc descriptors. An exception was the melting points, which were not well modeled by any of the available indices. Factor analysis (principal component analysis) was used to examine the intrinsic dimensionalities of the data and parameter sets. A single factor accounts for about 82% of the variance in the eight physical properties, two factors account for 94%, and three factors account for about 99%. The melting points load strongly on a factor independent of the other properties. Of the examined parameter sets, the connectivity indices exhibited the highest dimensionality.

It is obvious that the structure of a compound, both geometric and electronic, determines its properties. Nonetheless, elucidation of the connection between molecular structure and bulk properties has posed a challenge to chemists for more than a century. A central issue has been how to quantify the elusive concept of "structure". In early attempts properties were commonly represented as sums of contributions from atoms, bonds, or larger structural subunits.[1-7] These approaches often required a large number of empirical parameters, and recently attention has turned to the use of more general structural parameters, in particular those derived from chemical graph theory.[8-15] Among the most successful of these general parameters have been Wiener distance indices,[16-19] connectivity indices,[9,13-15,20] information indices,[12,21]

and ad hoc descriptors.[15,22] In some cases molecular volumes and surface areas have also been suggested as important variables, most notably for solubility properties.[23-29]

The alkanes represent an especially attractive class of compounds as a starting point for the application of molecular modeling techniques. Many properties of the alkanes vary in a regular manner with molecular mass and extent of branching, and because the alkanes are nonpolar, a number of complexities that arise with more polar compounds are avoided. Our purposes in the present report are threefold. First, we hope to obtain practical structure–property equations for eight representative physical properties of the alkanes, utilizing relatively simple structural parameters. Such equations can be used to predict values for as yet unmeasured properties of compounds and also, in some cases, may aid in the "design" of compounds with properties suitable for special purposes. Second, we wish to evaluate the relative performances of the above descriptor sets in relating the alkane molecular structures to the set of physical properties. Third, we hope to draw inferences from our results concerning the natures of the physical forces responsible for the properties observed. As an aid in this last effort, we have utilized the technique of factor analysis to estimate the inherent dimensionality of the set of physical properties investigated and also the dimensionalities of the parameter sets.

Methods

Data. Eight representative physical properties were selected, based in part on their perceived importances and in part on the availability of a suitable body of data: boiling points (bp), molar volumes at 20 °C (MV), molar refractions at 20 °C (MR), heats of vaporization at 25 °C

(1) Kopp, H. *Justus Liebigs Ann. Chem.* **1842**, *41*, 79, 169. Bruhl, J. W. Z. *Phys. Chem., Stoechiom. Verwandtschaftsl.* **1891**, *7*, 140.

(2) Kinney, C. R. *J. Am. Chem. Soc.* **1938**, *60*, 3032. Exner, O. *Collect. Czech. Chem. Commun.* **1967**, *32*, 1.

(3) Denbigh, K. G. *Trans. Faraday Soc.* **1940**, *36*, 936. Vogel, A. I. *J. Chem. Soc.* **1948**, 1833. Miller, K. J.; Savchic, J. A. *J. Am. Chem. Soc.* **1979**, *101*, 7206.

(4) Franklin, J. L. *Ind. Eng. Chem.* **1949**, *41*, 1070.

(5) Somayajulu, G. R.; Zwolinski, B. J. *J. Phys. Chem.* **1966**, *70*, 3498; *Trans. Faraday Soc.* **1972**, *68*, 1971.

(6) Benson, S. W.; Cruickshank, F. R.; Golden, D. M.; et al. *Chem. Rev.* **1969**, *69*, 279. Rouvray, D. H. *CHEMTECH* **1973**, *3*, 379.

(7) Lyman, W. J.; Reehl, W. F.; Rosenblatt, D. H. *Handbook of Chemical Property Estimation Methods*; McGraw-Hill: New York, 1982.

(8) Balaban, A. T., Ed. *Chemical Applications of Graph Theory*; Academic: New York, 1976.

(9) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic: New York, 1976.

(10) Trinajstic, N. *Chemical Graph Theory*; CRC: Boca Raton, FL, 1983; two volumes.

(11) King, R. B., Ed. *Chemical Applications of Topology and Graph Theory*; Elsevier: Amsterdam, The Netherlands, 1983.

(12) Bonchev, D. *Information Theoretic Indices for Characterization of Chemical Structure*; Wiley-Interscience: New York, 1983.

(13) Rouvray, D. H. *Sci. Am.* **1986**, *254*(9), 40.

(14) Kier, L. b.; Hall, L. H. *Molecular Connectivity in Structure-Activity Analysis*; Wiley: New York, 1986.

(15) Seybold, P. G.; May, M. A.; Bagal, U. A. *J. Chem. Educ.* **1987**, *64*, 575.

(16) Wiener, H. *J. Am. Chem. Soc.* **1947**, *69*, 17, 2636.

(17) Wiener, H. *J. Chem. Phys.* **1947**, *15*, 766.

(18) Wiener, H. *J. Phys. Chem.* **1948**, *52*, 425, 1082.

(19) Platt, J. R. *J. Phys. Chem.* **1952**, *56*, 328.

(20) Randić, M. *J. Am. Chem. Soc.* **1975**, *97*, 6609.

(21) Basak, S. C.; Harriss, D. K.; Magnuson, V. R. *J. Pharm. Sci.* **1984**, *73*, 429.

(22) Seybold, P. G.; May, M. A.; Gargas, M. L. *Acta Pharm. Jugosl.* **1986**, *36*, 253.

(23) Bondi, A. *J. Phys. Chem.* **1964**, *68*, 441.

(24) Hermann, R. B. *J. Phys. Chem.* **1972**, *76*, 2754.

(25) McAuliffe, C. *J. Phys. Chem.* **1966**, *70*, 1267.

(26) Moriguchi, I.; Kanada, Y. *Chem. Pharm. Bull.* **1977**, *25*, 926.

(27) Mellors, A.; McGowan, J. *Biochem. Pharm.* **1985**, *34*, 2413.

(28) Charton, M. *Top. Curr. Chem.* **1983**, *114*, 107.

(29) Kamlet, M. J.; Doherty, R. M.; Abboud, J.-L. M.; Abraham, M. H.; Taft, R. W. *CHEMTECH* **1986**, *16*, 566.

Molecular Modeling of Alkanes

J. Am. Chem. Soc., Vol. 110, No. 13, 1988 4187

(HV), critical temperatures (TC), critical pressures (PC), surface tensions at 20 °C (ST), and melting points (mp). Values for most properties were taken from ref 30, but some values were obtained from other sources.[31-34] Molar volumes were calculated as MW/d, where MW is the molecular weight and d is the density (g/cm^3) at 20 °C. Molar refractions were calculated from the index of refraction n_0, MV, and d, using the Lorentz–Lorenz expression

$$MR = \frac{n_0^2 - 1}{n_0^2 + 2} \frac{MW}{d}$$

MR is of interest because of its relationship to molecular polarizability.[3,30,35,36] Values of the properties employed for the 74 C_2–C_9 alkanes are shown in Table I. Only liquid-phase values for MV, MR, HV, and ST were used. The extremely hindered compound 2,2,3,3-tetramethylbutane (39) is a solid at 25 °C; its mp (+100.7 °C) was judged to be an outlier. Compounds 1–4 are gases at 20 °C and 1 atm.

Parameters. Four decades ago Wiener introduced an index based on path distances.[16,17] The *Wiener index w* is the sum of all unique, shortest path distances, in terms of C–C bonds, in the carbon skeleton ("hydrogen-suppressed graph") of the hydrocarbon. For example, for 2-methylbutane

$$\begin{array}{c} C_5 \\ | \\ C_1 - C_2 - C_3 - C_4 \end{array}$$
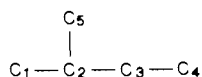
the Wiener index is obtained as half the sum of the elements in the *distance matrix* **D**

$$\mathbf{D} = \begin{pmatrix} 0 & 1 & 2 & 3 & 2 \\ 1 & 0 & 1 & 2 & 1 \\ 2 & 1 & 0 & 1 & 2 \\ 3 & 2 & 1 & 0 & 3 \\ 2 & 1 & 2 & 3 & 0 \end{pmatrix}$$

i.e., $w = 18$. (In the distance matrix each unique path is included twice, as d_{ij} and as d_{ji}.) The Wiener index increases with the number of carbon atoms; it is lower for branched isomers than for more extended isomers. In many applications the *reduced Wiener index*, $w_r = w/N_c^2$, is used, where N_c is the number of carbon atoms.[15] Wiener also introduced the parameter p_3, the number of unique paths of length 3 C–C bonds, originally as a "polarizability" factor[16,17] but later associated with steric crowding.[19] As seen from either the carbon skeleton or the distance matrix, for 2-methylbutane, $p_3 = 2$. In some cases a *modified Wiener index*, $w_{mod} = w + p_3$, has been used. Platt[19] introduced the parameter f, calculated by taking the number of adjacent bonds to each bond and summing these values for all bonds in the carbon skeleton. For 2-methylbutane $f = 2(C_1-C_2) + 3(C_2-C_3) + 1(C_3-C_4) + 2(C_2-C_5) = 8$. The inverses of w and w_r were also included in this parameter set.

Randić introduced the concept of a connectivity (branching) index in 1975.[20] The concept has been further developed and applied extensively by Randić, Kier, Hall, and their coworkers.[9,14] Each carbon atom i in the carbon skeleton of a hydrocarbon is assigned a *valence*, δ_i, equal to the number of carbon bonds to that atom. The zeroth-order connectivity index $^0\chi$ is then the sum over all *atoms* i (eq 1). For 2-methylbutane

$$^0\chi = \sum_{\text{atoms}} 1/\delta_i^{1/2} \qquad (1)$$

$^0\chi = 1/\sqrt{1} + 1/\sqrt{3} + 1/\sqrt{2} + 1/\sqrt{1} + 1/\sqrt{1} = 4.284$. The first-order connectivity index $^1\chi$ is the sum over all *bonds* i–j (eq 2). For

$$^1\chi = \sum_{\text{bonds}} 1/(\delta_i\delta_j)^{1/2} \qquad (2)$$

2-methylbutane $^1\chi = 1/\sqrt{2} + 1/\sqrt{6} + 1/\sqrt{2} + 1/\sqrt{3} = 2.400$. The second-order index $^2\chi$ is a sum over 2-bond *paths i–j–k* of terms $1/(\delta_i\delta_j\delta_k)^{1/2}$. A hierarchy of additional indices $^m\chi_t$ of order m and type t can be obtained by summing analogous terms over substructural units involving paths (t = p), clusters (t = c), or path–cluster (t = pc) combinations of m bonds. Values for these indices for the 74 alkanes were

taken from Appendix I of Kier and Hall's book.[9] The total structure index χ_t can be calculated with all the carbon valences of the compound (eq 3). The above indices, plus the inverses of $^0\chi$, $^1\chi$, and χ_t, were used in this parameter set.

$$\chi_t = 1/(\delta_1 \cdot \delta_2 \cdots \delta_m)^{1/2} \qquad (3)$$

Information indices have been used successfully in a variety of applications in chemistry.[12,21] Several different types of information indices can be defined.[12,38] When Kier's approach[39] is followed, the atoms of a compound are divided into classes according to their equivalence or nonequivalence. In 2-methylbutane, carbons 1 and 5 are equivalent and the remaining carbons are unique; therefore, there are four classes of carbon atoms, one containing two members and three containing one member each. Likewise, there are four classes of hydrogen atoms containing six, one, two, and three members, respectively. Following from Shannon's information theorem, Kier's information index ("molecular negentropy") I is defined as[39] in eq 4 where $p_i = n_i/N$ is the probability

$$\begin{aligned} I &= -N\sum p_i \log p_i \\ &= N \log N - \sum n_i \log n_i \end{aligned} \qquad (4)$$

of class i, n_i is the number of members in class i, and N is the total number of atoms in the compound. If only carbon atoms are considered, one obtains the index ICM. Basak[21] has defined additional indices, as

$$IC = I/N \qquad (5)$$

$$SIC = IC/\log N$$

$$CIC = (1/N)\sum n_i \log n_i$$

The indices I, ICM, IC, SIC, and CIC together formed the information index parameter set.

In some recent studies specific ad hoc indices have been shown to give a good account of several properties.[15,22,40] The simplest representative of molecular mass or volume is N_c, the number of carbon atoms. Because each branch of a hydrocarbon terminates as a methyl group, the number of terminal methyl groups, T_m, is a crude measure of branching. Randić first suggested the use of T_3, the number of terminal methyls separated by 3 bonds, as a steric parameter in analysis of chromatographic retention data.[41] The steric index p_3 has been identified above in connection with the Wiener scheme. The inverse $1/N_c$ and the quadratic terms N_c^2, T_3^2, and T_m^2 were also included in this parameter set.

Molecular volumes V_B and surface areas A_B were calculated from sums of contributions as described by Bondi.[23] Values for the most significant parameters described above for the 74 alkanes are given in Table II.

Regression equations and other statistical measures were obtained by options in the SAS software package[42] on an IBM 8083E computer.

Factor Analysis. Factor analysis[43-46] is a multivariate statistical technique for estimating the inherent dimensionality of a set of variables. It rests on the assumption that the experimental variables of interest can be considered to depend on a (presumably smaller) set of hypothetical underlying variables or factors. Observed correlations among the experimental variables are assumed to result from their common dependences on these underlying factors.

In principal component analysis (PCA) a matrix is first constructed consisting of the correlations (or covariances) among the variables of interest. The eigenvalues and eigenvectors of this matrix are then determined. The eigenvectors so obtained are orthogonal, and the sum of their eigenvalues equals the original number of variables. Each eigenvector is a linear combination of the original variables and represents a principal component, or initial factor. The coefficient of an observed variable in a factor is termed the *loading* of that variable on the factor. The relative importance of each factor is related to the magnitude of its eigenvalue. The process can be viewed as one in which the first principal component axis is constructed to account for a maximum amount of variance in the data; the second component axis accounts for a maximum amount of the remaining variance under the constraint that it be or-

(30) Driesbach, R. R., Ed. *Physical Properties of Chemical Compounds*; Advances in Chemistry 22; American Chemical Society: Washington, DC, 1959; Vol. II.

(31) Weast, R. C., Ed. *CRC Handbook of Chemistry and Physics*, 67th ed.; CRC: Boca Raton, FL, 1986–1987.

(32) Doss, M. P. *Physical Constants of the Principal Hydrocarbons*, 3rd ed.; The Texas Co.: New York, 1942.

(33) Kobe, K. A.; Lynn, R. E., Jr. *Chem. Rev.* **1953**, *52*, 117.

(34) *Selected Values of Properties of Hydrocarbons and Related Compounds*; Thermodynamics Research Center, Texas A&M Research Center: College Station, TX, 1985.

(35) Pauling, L.; Pressman, D. *J. Am. Chem. Soc.* **1945**, *67*, 1003.

(36) Dunn, W. J., III *Eur. J. Med. Chem.-Chim. Ther.* **1977**, *12*, 109.

(37) Sabljic, A. *J. Chromatogr.* **1984**, *314*, 1.

(38) Balaban, A. T.; Motoc, I.; Bonchev, D.; Mekenyan, O. *Top. Curr. Chem.* **1983**, *114*, 21.

(39) Kier, L. B. *J. Pharm. Sci.* **1980**, *69*, 807.

(40) Hinckley, D. A.; Seybold, P. G.; Borris, D. P. *Spectrochim. Acta, Part A* **1986**, *42A*, 747.

(41) Randić, M. *J. Chromatogr.* **1978**, *161*, 1.

(42) SAS Institute, Inc., Box 8000, Cary, NC 27511-8000.

(43) Lawley, D. N.; Maxwell, A. E. *Factor Analysis as a Statistical Method*; Butterworths: London, 1963.

(44) Kim. J.-O.; Mueller, C. W. *Factor Analysis: Statistical Methods and Practical Issues*; Sage: Beverly Hills, CA, 1978.

(45) Malinowski, E. R.; Howery, D. G. *Factor Analysis in Chemistry*; Wiley-Interscience: New York, 1980.

(46) *SAS User's Guide; Statistics*; SAS Institute: Cary, NC, 1985.

Table I. Experimental Values for the Physical Properties of the 74 Alkanes

observn	name	bp	MV	MR	HV	TC	PC	ST	mp
1	2	−88.630				32.27	48.20		−183.27
2	3	−42.070				96.80	42.01		−187.69
3	4	−0.500				152.01	37.47		−138.35
4	2M3	−11.730				134.98	36.00		−159.60
5	5	36.074	115.205	25.2656	26.42	196.62	33.31	16.00	−129.72
6	2M4	27.852	116.426	25.2923	24.59	187.80	32.90	15.00	−159.90
7	22MM3	9.503	122.074	25.7243	21.78	160.60	31.57		−16.55
8	6	68.740	130.688	29.9066	31.55	234.70	29.92	18.42	−95.35
9	2M5	60.271	131.933	29.9459	29.86	224.90	29.95	17.38	−153.67
10	3M5	63.282	129.717	29.8016	30.27	231.20	30.83	18.12	−118.00
11	22MM4	49.741	132.744	29.9347	27.69	216.20	30.67	16.30	−99.87
12	23MM4	57.988	130.240	29.8104	29.12	227.10	30.99	17.37	−128.54
13	7	98.427	146.540	34.5504	36.55	267.01	27.01	20.26	−90.61
14	2M6	90.052	147.656	34.5908	34.80	257.90	27.20	19.29	−118.28
15	3M6	91.850	145.821	34.4597	35.08	262.40	28.10	19.79	−119.40
16	3E5	93.475	143.517	34.2827	35.22	267.60	28.60	20.44	−118.60
17	22MM5	79.197	148.695	34.6166	32.43	247.70	28.40	18.02	−123.81
18	23MM5	89.784	144.153	34.3237	34.24	264.60	29.20	19.96	−119.10
19	24MM5	80.500	148.949	34.6192	32.88	247.10	27.40	18.15	−119.24
20	33MM5	86.064	144.530	34.3323	33.02	263.00	30.00	19.59	−134.46
21	223MMM4	80.882	145.191	34.3736	32.04	258.30	29.75	18.76	−24.91
22	8	125.665	162.592	39.1922	41.48	296.20	24.64	21.76	−56.79
23	2M7	117.647	163.663	39.2316	39.68	288.00	24.80	20.60	−109.04
24	3M7	118.925	161.832	39.1001	39.83	292.00	25.60	21.17	−120.50
25	4M7	117.709	162.105	39.1174	39.67	290.00	25.60	21.00	−120.95
26	3E6	118.534	160.072	38.9441	39.40	292.00	25.74	21.51	
27	22MM6	106.840	164.285	39.2525	37.29	279.00	25.60	19.60	−121.18
28	23MM6	115.607	160.395	38.9808	38.79	293.00	26.60	20.99	
29	24MM6	109.429	163.093	39.1300	37.76	282.00	25.80	20.05	−137.50
30	25MM6	109.103	164.697	39.2596	37.86	279.00	25.00	19.73	−91.20
31	33MM6	111.969	160.879	39.0087	37.93	290.84	27.20	20.63	−126.10
32	34MM6	117.725	158.814	38.8453	39.02	298.00	27.40	21.64	
33	23ME5	115.650	158.794	38.8362	38.52	295.00	27.40	21.52	−114.96
34	33ME5	118.259	157.026	38.7171	37.99	305.00	28.90	21.99	−90.87
35	223MMM5	109.841	159.526	38.9249	36.91	294.00	28.20	20.67	−112.27
36	224MMM5	99.238	165.083	39.2617	35.13	271.15	25.50	18.77	−107.38
37	233MMM5	114.760	157.292	38.7617	37.22	303.00	29.00	21.56	−100.70
38	234MMM5	113.467	158.852	38.8681	37.61	295.00	27.60	21.14	−109.21
39	2233MMMM4	106.470				270.80	24.50		
40	9	150.798	178.713	43.8423	46.44	322.00	22.74	22.92	−53.52
41	2M8	143.260	179.773	43.8795	44.65	315.00	23.60	21.88	−80.40
42	3M8	144.180	177.952	43.7296	44.75	318.00	23.70	22.34	−107.60
43	4M8	142.480	178.150	43.7687	44.75	318.30	23.06	22.34	−113.20
44	3E7	143.000	176.410	43.6420	44.81	318.00	23.98	22.81	−114.90
45	4E7	141.200	175.685	43.4907	44.81	318.30	23.98	22.81	
46	22MM7	132.690	180.507	43.9138	42.28	302.00	22.80	20.80	−113.00
47	23MM7	140.500	176.653	43.6269	43.79	315.00	23.79	22.34	−116.00
48	24MM7	133.500	179.120	43.7393	42.87	306.00	22.70	21.30	
49	25MM7	136.000	179.371	43.8484	42.87	307.80	22.70	21.30	
50	26MM7	135.210	180.914	43.9258	42.82	306.00	23.70	20.83	−102.90
51	33MM7	137.300	176.897	43.6870	42.66	314.00	24.19	22.01	
52	34MM7	140.600	175.349	43.5473	43.84	322.70	24.77	22.80	
53	35MM7	136.000	177.386	43.6378	42.98	312.30	23.59	21.77	
54	44MM7	135.200	176.897	43.6022	42.66	317.80	24.18	22.01	
55	23ME6	138.000	175.445	43.6550	43.84	322.70	24.77	22.80	
56	24ME6	133.800	177.386	43.6472	42.98	330.3	25.56	21.77	
57	33ME6	140.600	173.077	43.2680	43.04	327.2	25.66	23.22	
58	34ME6	140.400	172.844	43.3746	43.95	312.3	23.59	23.27	
59	123MMM6	133.600	175.878	43.6226	41.91	318.1	25.07	21.86	
60	224MMM6	126.540	179.220	43.7638	40.57	301.0	23.39	20.51	−120.00
61	225MMM6	124.084	181.346	43.9356	40.17	296.6	22.41	20.04	−105.78
62	233MMM6	137.680	173.780	43.4347	42.23	326.1	25.56	22.41	−116.80
63	234MMM6	139.000	173.498	43.3917	42.93	324.2	25.46	22.80	
64	235MMM6	131.340	177.656	43.6474	41.42	309.4	23.49	21.27	−127.80
65	244MMM6	130.648	177.187	43.6598	40.84	309.1	23.79	21.17	−113.38
66	334MMM6	140.460	172.055	43.3407	42.28	330.6	26.45	23.27	−101.20
67	33EE5	146.168	170.185	43.1134	43.36	342.8	26.94	23.75	−33.11
68	223MME5	133.830	174.537	43.4571	42.02	322.6	25.96	22.38	−99.20
69	233MME5	142.000	170.093	42.9542	42.55	338.6	26.94	23.87	
70	234MEM5	136.730	173.804	43.4037	42.93	324.2	25.46	22.80	−122.20
71	2233(M)5	140.274	169.495	43.2147	41.00	334.5	27.04	23.38	−9.90
72	2234(M)5	133.016	173.557	43.4359	41.00	319.6	25.66	21.98	−121.09
73	2244(M)5	122.284	178.256	43.8747	38.10	301.6	24.58	20.37	−66.54
74	2334(M)5	141.551	169.928	43.2016	41.75	334.5	26.85	23.31	−102.12

[a] Units are the following: bp, °C; MV, cm³/mol at 20 °C; MR, cm³/mol; HV, kJ/mol; TC, °C; PC, atm; ST, dyn/cm; mp, °C. [b] Abbreviations as in Kier and Hall:[9] 2 = ethane, 3 = propane, etc.; M = methyl, E = ethyl; e.g., 34ME6 = 3-methyl-4-ethylhexane.

Table II. Selected Parameter Values for the 74 Alkanes

observn	name	W	p_3	f	$^0\chi$	$^1\chi$	χ_t	N_c	T_m	T_3	I
1	2	1	0	0	2.000 00	1.000 00	1.0000	2	2	0	1.9538
2	3	4	0	2	2.707 10	1.414 21	0.7071	3	2	0	5.5823
3	4	10	1	4	3.414 21	1.914 21	0.5000	4	2	1	7.7645
4	2M3	9	0	6	3.577 35	1.732 05	0.5774	4	3	0	6.0262
5	5	20	2	6	4.121 32	2.414 21	0.3536	5	2	0	12.0343
6	2M4	18	2	8	4.284 45	2.270 05	0.4082	5	3	2	13.6132
7	22MM3	16	0	12	4.500 00	2.000 00	0.5000	5	4	0	5.5592
8	6	35	3	8	4.828 42	2.914 21	0.2500	6	2	0	14.7290
9	2M5	32	3	10	4.991 56	2.770 05	0.2887	6	3	0	18.1141
10	3M5	31	4	10	4.991 56	2.808 06	0.2887	6	3	2	16.3080
11	22MM4	28	3	14	5.207 10	2.560 66	0.3536	6	4	3	13.9676
12	23MM4	29	4	12	5.154 70	2.642 73	0.3333	6	4	4	9.4581
13	7	56	4	10	5.535 53	3.414 21	0.1768	7	2	0	19.4261
14	2M6	52	4	12	5.698 67	3.270 05	0.2041	7	3	0	22.8112
15	3M6	50	5	12	5.698 67	3.308 06	0.2041	7	3	1	25.2195
16	3E5	48	6	12	5.698 67	3.346 06	0.2041	7	3	0	15.1999
17	22MM5	46	4	16	5.914 21	3.060 66	0.2500	7	4	0	18.6647
18	23MM5	46	6	14	5.861 80	3.180 73	0.2357	7	4	3	22.5840
19	24MM5	48	4	14	5.861 80	3.125 89	0.2357	7	4	0	14.1551
20	33MM5	44	6	16	5.914 21	3.121 32	0.2500	7	4	4	17.7675
21	223MMM4	42	6	18	6.077 35	2.943 37	0.2887	7	5	6	16.0292
22	8	84	5	12	6.242 64	3.914 21	0.1250	8	2	0	22.4874
23	2M7	79	5	14	6.405 77	3.770 05	0.1443	8	3	0	27.6787
24	3M7	76	6	14	6.405 77	3.808 06	0.1443	8	3	1	30.0870
25	4M7	75	6	14	6.405 77	3.808 06	0.1443	8	3	0	24.0664
26	3E6	72	7	14	6.405 77	3.846 06	0.1443	8	3	0	25.8726
27	22MM6	71	5	18	6.621 32	3.560 66	0.1768	8	4	0	23.5322
28	23MM6	70	7	16	6.568 91	3.680 73	0.1667	8	4	2	27.4515
29	24MM6	71	6	16	6.568 91	3.663 90	0.1667	8	4	1	27.4515
30	25MM6	74	5	16	6.568 91	3.625 89	0.1667	8	4	0	17.2165
31	33MM6	67	7	18	6.621 32	3.621 32	0.1768	8	4	2	26.8494
32	34MM6	68	8	16	6.568 91	3.718 74	0.1667	8	4	3	22.0330
33	23ME5	67	8	16	6.568 91	3.718 74	0.1667	8	4	0	23.2371
34	33ME5	64	9	18	6.621 32	3.681 98	0.1768	8	4	3	19.2381
35	223MMM5	63	8	20	6.784 45	3.481 38	0.2041	8	5	4	23.3050
36	224MMM5	66	5	20	6.784 45	3.416 50	0.2041	8	5	0	20.8967
37	233MMM5	62	9	20	6.784 45	3.504 03	0.2041	8	5	6	24.2139
38	234 MMM5	65	8	18	6.732 05	3.553 41	0.1925	8	5	4	18.7954
39	2233MMMM4	58	9	24	7.000 00	3.250 00	0.2500	8	6	9	8.9234
40	9	120	6	14	6.949 74	4.414 21	0.0884	9	2	0	27.5056
41	2M8	114	6	16	7.112 88	4.270 05	0.1021	9	3	0	32.6969
42	3M8	110	7	16	7.112 88	4.308 06	0.1021	9	3	1	35.1052
43	4M8	108	7	16	7.112 88	4.308 06	0.1021	9	3	0	35.1052
44	3E7	104	8	16	7.112 88	4.346 06	0.1021	9	3	0	30.8907
45	4E7	102	8	16	7.112 88	4.346 06	0.1021	9	3	0	29.0846
46	22MM7	104	6	20	7.328 42	4.060 66	0.1250	9	4	0	28.5504
47	23MM7	102	8	18	7.276 02	4.180 73	0.1179	9	4	2	32.4697
48	24MM7	102	7	18	7.276 02	4.163 90	0.1179	9	4	0	32.4697
49	25MM7	104	7	18	7.276 02	4.163 90	0.1179	9	4	1	32.4697
50	26MM7	108	6	18	7.276 02	4.125 89	0.1179	9	4	0	22.2346
51	33MM7	98	8	20	7.328 42	4.121 32	0.1250	9	4	2	31.8676
52	34MM7	98	9	18	7.276 02	4.218 74	0.1179	9	4	2	34.8779
53	35MM7	100	8	18	7.276 02	4.201 90	0.1179	9	4	2	27.0511
54	44MM7	96	8	20	7.328 42	4.121 32	0.1250	9	4	0	25.8470
55	23ME6	96	9	18	7.276 02	4.218 74	0.1179	9	4	0	32.4697
56	24ME6	98	8	18	7.276 02	4.201 90	0.1179	9	4	0	28.2552
57	33ME6	92	10	20	7.328 42	4.181 98	0.1250	9	4	2	30.0614
58	34ME6	94	10	18	7.276 02	4.256 74	0.1179	9	4	1	30.6635
59	223MMM6	92	9	22	7.491 56	3.981 38	0.1443	9	5	3	28.3231
60	224MMM6	94	7	22	7.491 56	3.954 50	0.1443	9	5	1	28.3231
61	225MMM6	98	6	22	7.491 56	3.916 50	0.1443	9	5	0	25.9149
62	233MMM6	90	10	22	7.491 56	4.004 03	0.1443	9	5	4	29.2321
63	234MMM6	92	10	20	7.439 15	4.091 42	0.1361	9	5	4	32.2424
64	235MMM6	96	8	20	7.439 15	4.036 58	0.1361	9	5	2	29.8342
65	244MMM6	92	8	22	7.491 56	3.977 16	0.1443	9	5	2	29.2321
66	334MMM6	88	11	22	7.491 56	4.042 04	0.1443	9	5	5	31.6404
67	33EE5	88	12	20	7.328 42	4.242 64	0.1250	9	4	0	17.4182
68	223MME5	88	10	22	7.491 56	4.019 38	0.1443	9	5	0	24.1087
69	233MME5	86	12	22	7.491 56	4.064 69	0.1443	9	5	4	27.4259
70	234MEM5	90	10	20	7.439 15	4.091 42	0.1361	9	5	0	23.8136
71	2233MMMM	82	12	26	7.707 10	3.810 66	0.1768	9	6	8	25.0856
72	2234MMMM	86	10	24	7.654 70	3.854 05	0.1667	9	6	5	25.6877
73	2244MMMM	88	6	26	7.707 10	3.707 10	0.1768	9	6	0	13.9416
74	2334MMMM	84	12	24	7.654 70	3.886 75	0.1667	9	6	8	20.5760

thogonal to the first component, and so forth, until all component axes are constructed. For every object in the data set a numerical value for each factor can be obtained from the values of the variables; these numerical values are termed factor *scores*. PCA has previously been applied

4190 *J. Am. Chem. Soc., Vol. 110, No. 13, 1988*

Needham et al.

Table III. Correlations among the Physical Properties Examined

	bp	MV	MR	HV	TC	PC	ST	mp
bp	1.000							
MV	0.948	1.000						
MR	0.972	0.992	1.000					
HV	0.989	0.942	0.957	1.000				
TC	0.993	0.911	0.952	0.955	1.000			
PC	−0.932	−0.931	−0.892	−0.875	−0.898	1.000		
ST	0.946	0.807	0.865	0.922	0.965	−0.650	1.000	
mp	0.457	0.143	0.157	0.122	0.472	−0.436	0.421	1.000

to liquid-state properties,[47-49] structural descriptors,[50] and other chemical problems.[51,52]

The question of how many components are significant in such an analysis is not simply answered. In the social sciences a common guide has been to retain only components with eigenvalues greater than unity, but this requirement is surely too stringent for the more accurate physical data analyzed here; in general, it is deemed wisest to use several criteria to assist in making this decision.[44] Possible criteria include a lower acceptable magnitude for the retained eigenvalues (based on the perceived accuracy of the data), the scree-test, and Kaiser's measure of sampling adequacy, MSA.[44,46] Despite these aids, any decision on this question remains somewhat subjective.

Principal component analysis does not necessarily lead to the most physically meaningful or readily interpretable factorization. First, it is often possible, with rather standard statistical criteria, to achieve a "simpler" factor structure by performing orthogonal rotations on the initial factors. This arises because neither the factors nor their loadings are uniquely defined. Second, in the present example there is no a priori physical reason to require that the factors be orthogonal; oblique rotations can be expected in some cases to yield more reasonable representations of the factor structure. Subsequent rotations do not, however, change the number of factors or the total fraction of variance accounted for by the factors.

Factor analysis, including PCA and orthogonal and oblique rotations, was performed with the VARIMAX and PROMAX options within PROC FACTOR of the SAS software package.[42,46]

Results

Before other analyses of the data are attempted, it is instructive to examine the correlations among the physical properties for the 74 alkanes, as shown in Table III. It is readily apparent that strong internal correlations are present among most of the properties. Melting points are an exception, being only weakly correlated with the other properties. Surface tension is another possible exception. For the remaining six properties all the correlation coefficients (r) are greater than 0.875, and for the subset (bp, MV, MR, HV), all correlations exceed 0.942. It may be anticipated, then, that the factor analysis should reveal some lower (than eight) number of inherent dimensions for these physical properties. Furthermore, it can be anticipated that if a given set of structural parameters successfully models a given property, this parameter set should also be reasonably successful in modeling other, strongly correlated properties. The converse can also be anticipated; i.e., lack of success should transfer to correlated properties.

Regression Equations. Table IV shows coefficients of determination (r^2) and standard errors (s) for the most successful one-, two-, and three-parameter models of the eight physical properties. (Note that r^2 gives the fraction of the variance in the data that is accounted for by the model.) Overall, the most successful single parameters are seen to be the connectivity indices $^1\chi$ and $^0\chi$, the number of carbons N_c, and the calculated (Bondi) volume V_B. Correlations among these parameters and some other selected
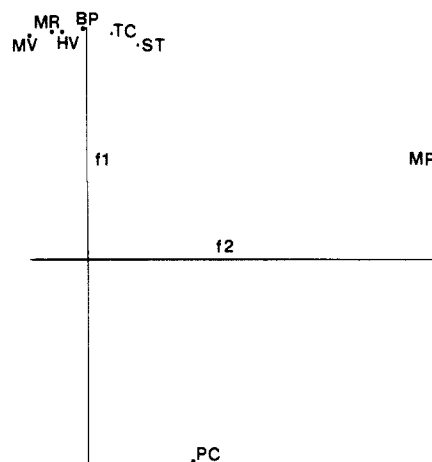
(47) Cramer, R. D., III *J. Am. Chem. Soc.* **1980**, *102*, 1837, 1849; *Quant. Struct.-Act. Relat. Pharmacol., Chem. Biol.* **1983**, *2*, 7.

(48) Chastrette, M.; Rajzmann, M.; Chanon, M.; Purcell, K. F. *J. Am. Chem. Soc.* **1985**, *107*, 1.

(49) Dunn, W. J., III; Wold, S. *Acta Chem. Scand., Ser. B* **1978**, *B32*, 536.

(50) Basak, S. C.; Magnuson, V. R.; Niemi, G. J.; Regal, R. R.; Veith, G. D. In *Mathematical Modelling in Science and Technology*; Avala, X. J. R., Leitman, G., Mote, C. D., Rodin, E. Y., Eds.; Proceedings 5th ICMM; Pergamon: New York, 1986.

(51) Weiner, P. H. *CHEMTECH* **1977**, *7*, 321.

(52) Hellberg, S.; Wold, S.; Dunn, W. J., III; Gasteiger, J.; Hutchings, M. G. *Quant. Struct.-Act. Relat. Pharmacol., Chem. Biol.* **1985**, *4*, 1.

Figure 1. Principal components $f1$ and $f2$ for the eight physical properties.

indices are shown in Table V. It is apparent that the leading single parameters are themselves intercorrelated and related to molecular volume.

Whereas the connectivity indices and ad hoc descriptors typically are superior to the Wiener and information indices at the single-term level, it is evident from Table IV that the latter parameter sets tend to "catch up" as additional terms are included. At the three-term level the four approaches tend toward a degree of similarity among their multiple correlation coefficients. On the basis of the standard errors s—which are more sensitive measures than r^2—overall, the connectivity indices and ad hoc descriptors are superior to the other indices in fitting the data for the present set of alkane properties.

Most of the physical properties are reasonably well accounted for by the present parameters even at the three-term level. The melting points (mp) are a notable exception, being not well modeled by any of the parameter sets. Surface tensions (ST) and critical pressures (PC) are less well accounted for generally than the remaining physical properties. As a rule, the fits of the properties at the three-term level fall in the same relative order as those from the single-term V_B and A_B models.

More exact regression equations for the eight physical properties are shown in Table VI for the connectivity indices and in Table VII for the ad hoc descriptors. With an aim of simplicity, these equations have been arbitrarily limited to no more than five parameters. Fewer parameters were included when F-statistics indicated that additional parameters were not significant. The ad hoc descriptors are as a rule more easily calculated and interpreted physically and thus may be more useful for some applications. For purposes of interpretation, the equations of Table VII have been dissected into mass (N_c, N_c^2, $1/N_c$), branching (T_m), and steric (T_3, p_3) contributions by multiplying the coefficients by the mean parameter values; fractional contributions were obtained by multiplying the absolute values by the coefficient of determination (r^2) and dividing by the sum. The results are shown in Table VIII. It is clear that molecular mass makes a dominant contribution for most properties, with smaller influences from branching and steric factors.

Factor Analysis. Results of the principal component analysis are shown in Table IX. In the analysis for all 8 properties and 51 compounds the first factor ($f1$) dominates and accounts for

Table IV. Summary of Results for One-, Two-, and Three-Parameter Models for the Eight Physical Properties

property	Wiener indices param	r^2	s	connectivity indices param	r^2	s	information indices param	r^2	s	ad hoc descriptors param	r^2	s	volume/area param	r^2	s
bp (°C)	w_{mod}	0.86	17.0	$^1\chi$	0.969	8.3	I	0.701	25.5	N_c	0.971	7.9	vol	0.971	7.9
$n = 74$	w_r p_3	0.982	6.3	$^1\chi$ $1/^0\chi$	0.991	4.4	I CIC	0.930	12.5	N_c N_c^2	0.982	6.4	area	0.962	9.1
	w_r p_3 $1/w_r$	0.995	3.4	$^1\chi$ χ_t $^4\chi_{pc}$	0.998	2.1	I CIC SIC	0.983	6.3	N_c N_c^2 T_m	0.990	4.7			
MV (cm³)	w_{mod}	0.953	3.8	$^0\chi$	0.925	4.8	I	0.622	10.7	N_c	0.974	2.8	vol	0.974	2.8
$n = 69$	w_r f	0.982	2.3	$^0\chi$ $^1\chi$	0.965	3.3	I CIC	0.964	3.3	N_c p_3	0.998	0.8	area	0.965	3.2
	w_r f p_3	0.983	2.3	$^0\chi$ $^1\chi$ $^3\chi_p$	0.996	1.1	I CIC SIC	0.975	2.8	N_c p_3 T_m	0.999	0.6			
MR (cm³)	w_{mod}	0.949	1.2	$^0\chi$	0.961	1.1	I	0.621	3.2	N_c	0.998	0.2	vol	0.999	0.2
$n = 69$	w_r f	0.974	0.9	$^0\chi$ $^1\chi$	0.995	0.4	I CIC	0.979	0.8	N_c p_3	0.9999	0.06	area	0.993	0.4
	w_r f p_3	0.990	0.5	$^0\chi$ $^1\chi$ $^3\chi_p$	0.999	0.1	I CIC SIC	0.998	0.2	N_c p_3 T_m	0.9999	0.05			
HV (kJ/mol)	w_{mod}	0.949	1.2	χ_t	0.939	1.3	I	0.704	2.9	N_c	0.918	1.6	vol	0.919	1.5
$n = 69$	w_r p_3	0.994	0.4	$^0\chi$ $^3\chi_c$	0.974	0.9	I CIC	0.917	1.6	N_c T_m	0.983	0.71	area	0.888	1.8
	w_r f p_3	0.994	0.4	$^0\chi$ $^3\chi_c$ $^4\chi_c$	0.992	0.5	I CIC ICM	0.971	0.9	N_c T_m p_3	0.993	0.46			
TC (°C)	w_{mod}	0.814	24.9	$^1\chi$	0.933	14.9	I	0.660	33.6	N_c	0.954	12.4	vol	0.954	12.4
$n = 74$	w_r p_3	0.965	10.9	$^1\chi$ $1/\chi_t$	0.982	7.8	I CIC	0.902	18.2	N_c N_c^2	0.969	10.2	area	0.949	13.0
	w_r p_3 $1/w_r$	0.990	5.9	$^1\chi$ $^4\chi_{pc}$ $1/^1\chi$	0.991	5.4	I CIC SIC	0.969	10.2	N_c $1/N_c$ p_3	0.984	7.4			
PC (atm)	w_r	0.869	1.6	χ_t	0.904	1.3	I	0.613	2.7	$1/N_c$	0.870	1.6	vol	0.863	1.6
$n = 74$	w_r f	0.930	1.2	$1/^0\chi$ $^5\chi_p$	0.955	0.9	IC CIC	0.914	1.3	N_c p_3	0.946	1.0	area	0.851	1.7
	w_r f $1/w_r$	0.957	0.9	$1/^0\chi$ $1/\chi_t$ $^3\chi_p$	0.967	0.8	IC CIC ICM	0.917	1.3	N_c p_3 $1/N_c$	0.974	0.7			
ST (dyne/cm)	p_3	0.774	0.9	$1/^1\chi$	0.832	0.8	I	0.490	1.4	$1/N_c$	0.775	0.9	vol	0.775	0.9
$n = 68$	w_r p_3	0.954	0.4	$1/^1\chi$ $^3\chi_p$	0.960	0.4	IC CIC	0.784	0.9	$1/N_c$ p_3	0.886	0.7	area	0.760	1.0
	w_r f p_3	0.971	0.3	$1/^1\chi$ $^3\chi_p$ $^2\chi$	0.970	0.3	IC CIC ICM	0.806	0.9	$1/N_c$ p_3 T_m	0.982	0.3			
mp (°C)	f	0.219	30.8	$1/^0\chi$	0.243	30.4	CIC	0.104	33.0	$1/N_c$	0.229	30.6	vol	0.199	31.2
$n = 56$				$1/^0\chi$ $^6\chi_c$	0.370	27.9	CIC IC	0.396	27.4	$1/N_c$ T_3^2	0.290	29.7	area	0.212	31.0
				$1/^0\chi$ $^6\chi_c$ $^4\chi_c$	0.407	27.4				$1/N_c$ T_3 T_3^2	0.367	28.3			

Table V. Correlations among Selected Parameters

	N_c	w_r	w_{mod}	$^0\chi$	$^1\chi$	χ_t	I	V_B
N_c	1.000							
w_r	0.892	1.000						
w_{mod}	0.955	0.914	1.000					
$^0\chi$	0.991	0.828	0.924	1.000				
$^1\chi$	0.978	0.942	0.968	0.941	1.000			
χ_t	−0.927	−0.921	−0.841	−0.892	−0.939	1.000		
I	0.832	0.850	0.865	0.784	0.878	−0.800	1.000	
V_B	1.000	0.892	0.995	0.990	0.978	−0.928	0.832	1.000

Table VI. Multiple Regression Equations for the Physical Properties Using Connectivity Indices

$$bp\ (°C) = -9.6\ (\pm4.1) + 38.1\ (\pm1.0){}^1\chi - 49.0\ (\pm19.3)1/{}^0\chi + 5.7\ (\pm0.3){}^4\chi_{pc} - 94.5\ (\pm9.8)\chi_t + 8.4\ (\pm2.5){}^6\chi_p$$
$$n = 74,\ r^2 = 0.999,\ s = 1.86,\ F = 9030$$

$$MV\ (cm^3) = 22.9\ (\pm0.7) + 14.6\ (\pm0.2){}^0\chi + 16.2\ (\pm0.4){}^1\chi - 7.3\ (\pm0.2){}^3\chi_p - 4.5\ (\pm0.3){}^4\chi_p - 2.1\ (\pm0.3){}^5\chi_c$$
$$n = 69,\ r^2 = 0.999,\ s = 0.5,\ F = 14\,294$$

$$MR\ (cm^3) = -0.8\ (\pm0.1) + 3.8\ (\pm0.02){}^0\chi + 4.6\ (\pm0.1){}^1\chi - 0.98\ (\pm0.03){}^3\chi_p - 0.63\ (\pm0.04){}^4\chi_p - 0.25\ (\pm0.06){}^5\chi_p$$
$$n = 69,\ r^2 = 0.9999,\ s = 0.05,\ F = 152\,558$$

$$HV\ (kJ/mol) = 1.9\ (\pm0.3) + 4.5\ (\pm0.2){}^0\chi + 4.8\ (\pm0.4){}^2\chi - 12.4\ (\pm0.5){}^3\chi_c + 20.7\ (\pm0.8){}^4\chi_c + 0.25\ (\pm0.05){}^4\chi_{pc}$$
$$n = 69,\ r^2 = 0.998,\ s = 0.2,\ F = 7849$$

$$TC\ (°C) = 150.7\ (\pm9.8) + 42.5\ (\pm2.4){}^1\chi + 12.8\ (\pm1.0){}^4\chi_{pc} - 159.7\ (\pm10.0)\chi_t - 11.5\ (\pm3.3){}^5\chi_p - 17.4\ (\pm3.4){}^6\chi_c$$
$$n = 74,\ r^2 = 0.995,\ s = 4.1,\ F = 2795$$

$$PC\ (atm) = 30.6\ (\pm0.6) + 17.5\ (\pm0.9)\chi_{t(k_t)} - 2.3\ (\pm0.2){}^2\chi + 1.9\ (\pm0.1){}^5\chi_{pc} - 2.6\ (\pm0.3){}^4\chi_p - 1.4\ (\pm0.4){}^6\chi_c$$
$$n = 74,\ r^2 = 0.981,\ s = 0.6,\ F = 694$$

$$ST\ (dyn/cm) = 31.3\ (\pm0.7) - 35.6\ (\pm1.3)1/{}^1\chi - 0.8\ (\pm0.1){}^2\chi + 0.8\ (\pm0.1){}^3\chi_p + 0.4\ (\pm0.1){}^5\chi_{pc} + 0.9\ (\pm0.1){}^5\chi_c$$
$$n = 68,\ r^2 = 0.986,\ s = 0.2,\ F = 845$$

$$mp\ (°C) = 39\ (\pm33) - 2192\ (\pm652)1/{}^0\chi + 859\ (\pm294)1/{}^1\chi - 94\ (\pm22){}^3\chi_c + 291\ (\pm62){}^4\chi_c + 77\ (\pm21){}^6\chi_c$$
$$n = 56,\ r^2 = 0.570,\ s = 23.8,\ F = 13$$

Table VII. Multiple Regression Equations for the Physical Properties Using ad Hoc Descriptors

$$bp\ (°C) = -152.9\ (\pm3.4) + 51.1\ (\pm1.0)N_c - 16.0\ (\pm1.5)T_m + 2.7(\pm0.2)p_3 - 1.8\ (\pm0.1)N_c{}^2 + 1.1\ (\pm0.2)T_m{}^2$$
$$n = 74,\ r^2 = 0.998,\ s = 2.0,\ F = 7786$$

$$MV\ (cm^3) = 6.8\ (\pm4.0) + 19.0\ (\pm0.3)N_c + 62.1\ (\pm13.5)1/N_c + 3.1\ (\pm0.3)T_m - 0.33\ (\pm0.04)T_m{}^2 - 1.81\ (\pm0.04)p_3$$
$$n = 69,\ r^2 = 0.999,\ s = 0.4,\ F = 20\,452$$

$$MR\ (cm^3) = 1.57\ (\pm0.05) + 4.78\ (\pm0.01)N_c - 0.14\ (\pm0.01)p_3 + 0.03\ (\pm0.01)T_m$$
$$n = 69,\ r^2 = 0.9999,\ s = 0.05,\ F = 267\,332$$

$$HV\ (kJ/mol) = 6.6\ (\pm0.5) + 4.7\ (\pm0.1)N_c - 1.95\ (\pm0.07)T_m + 0.28\ (\pm0.04)p_3 + 0.15\ (\pm0.04)T_3$$
$$n = 69,\ r^2 = 0.994,\ s = 0.4,\ F = 2746$$

$$TC\ (°C) = 77.5\ (\pm38.3) + 38.0\ (\pm7.5)N_c - 1.3\ (\pm0.4)N_c{}^2 + 6.3\ (\pm0.4)p_3 - 6.7\ (\pm0.7)T_m - 206.4\ (\pm55.8)1/N_c$$
$$n = 74,\ r^2 = 0.994,\ s = 4.8,\ F = 2074$$

$$PC\ (atm) = 40.7\ (\pm1.5) + 26.9\ (\pm3.2)1/N_c - 2.9\ (\pm0.2)N_c + 0.7\ (\pm0.1)p_3 - 0.02\ (\pm0.01)T_3{}^2 + 0.04\ (+0.01)T_m{}^2$$
$$n = 74,\ r^2 = 0.978,\ s = 0.7,\ F = 594$$

$$ST\ (dyn/cm) = 28.1\ (\pm0.5) - 50.1\ (\pm2.2)1/N_c - 1.7\ (\pm0.2)T_m + 0.11\ (\pm0.02)T_m{}^2 + 0.56\ (\pm0.02)p_3 + 0.05\ (\pm0.02)T_3$$
$$n = 68,\ r^2 = 0.989,\ s = 0.2,\ F = 1152$$

$$mp\ (°C) = -70.2\ (\pm10.5) - 248.8\ (\pm60.6)1/N_c - 12.6\ (\pm5.0)T_3 + 2.4\ (\pm0.8)T_3{}^2$$
$$n = 56,\ r^2 = 0.367,\ s = 28.3,\ F = 10$$

Table VIII. Estimates of the Relative Contributions of Molecular Mass, Branching, and Steric Factors to the Physical Properties[a]

property	mass	branching	steric	error
bp	+285.1 (0.82)	-44.5 (0.13)	+17.8 (0.05)	0.002
MV	+157.3 (0.89)	+6.7 (0.04)	-11.9 (0.07)	0.0006
MR	+37.5 (0.97)	+0.1 (0.003)	-0.9 (0.02)	0.0001
HV	+36.6 (0.79)	-7.6 (0.16)	+2.1 (0.04)	0.006
TC	+184.9 (0.73)	-26.1 (0.10)	+41.4 (0.16)	0.006
PC	-18.6 (0.77)	+0.6 (0.02)	+4.4 (0.18)	0.022
ST	-6.9 (0.44)	-4.8 (0.31)	+3.7 (0.24)	0.011
mp	-34.3[b] (0.34)		-2.8 (0.03)	0.633

[a] See text; fractional contributions shown in parentheses. [b] Dependence on $1/N_c$.

Table IX. Results of Principal Component Analysis for the Physical Properties: Eigenvalues and, in Parentheses, Cumulative Percent Variance Reproduced

factors	δ properties (51 compounds)	7 properties (68 compounds)
f1	6.598 (82.5)	6.442 (92.0)
f2	0.921 (94.0)	0.451 (98.5)
f3	0.378 (98.7)	0.088 (99.7)
f4	0.089 (99.8)	0.013 (99.9)
f5	0.011 (99.97)	0.004 (99.97)
f6	0.002 (99.99)	0.002 (99.99)
f7	0.0008 (100)	0.0004 (100)
f8	0.0003 (100)	

Table X. PCA Factor Loadings for the Physical Properties

property	PCA factor loadings		
	f1	f2	f3
bp	0.996	-0.011	0.079
MV	0.967	-0.149	-0.161
MR	0.983	-0.092	-0.058
HV	0.983	-0.065	0.042
TC	0.974	0.064	0.195
PC	-0.876	0.269	0.377
ST	0.925	0.131	0.343
mp	0.406	0.890	-0.207

82.5% of the variance in the properties examined. Addition of a second factor allows 94% of the property variance to be accounted for, and addition of a third factor allows almost 99% to be accounted for. When mp is excluded from the analysis (68 compounds), the first factor accounts for 92% of the variance in the remaining seven properties; a second factor allows over 98% of the variance in these seven properties to be reproduced.

Factor loadings from the PCA analysis are shown in Table X and Figure 1. These values represent correlations with the factors. With the exception of mp, the physical properties all are seen to load strongly on factor f1. Factor f2 is almost exclusively a melting point dimension; among the other properties only PC shows a significant loading on f2. The third factor f3 appears to be most significant for PC and ST and, to some extent, TC and mp. (In the seven-property (68 compounds) analysis with mp and hence f2 excluded, factors f1 and f3 retain virtually the same loadings as in the full analysis; therefore, this latter factor structure is not shown.)

In the present analysis of the physical properties further orthogonal and oblique rotations of the axes did not lead to a simplification of the factor structure. It is known that such rotations can often yield results that are not physically meaningful.[53]

It was of interest to attempt to model the factors arising from the factor analysis in order to assess their physical significance. f1 was strongly correlated ($r^2 = 0.968$) with the calculated molecular volumes, V_B. In terms of the ad hoc descriptors and connectivity indices, eq 6 and 7 were found.

$$f1 = -6.60\ (\pm0.08) + 0.92\ (\pm0.01)N_c - 0.20\ (\pm0.02)T_m + 0.007\ (\pm0.001)T_3{}^2\quad (6)$$
$$n = 51\qquad r^2 = 0.993\qquad s = 0.08$$

$$f1 = 2.8\ (\pm0.4) - 13.7\ (\pm0.9)1/{}^1\chi + 0.15\ (\pm0.02)1/\chi_t + 0.19\ (\pm0.02){}^4\chi_{pc}\quad (7)$$
$$n = 51\qquad r^2 = 0.988\qquad s = 0.1$$

(53) Swain, C. G.; Bryndza, H. E.; Swain, M. S. *J. Chem. Inf. Comput. Sci.* **1979,** *19,* 19.

Molecular Modeling of Alkanes

J. Am. Chem. Soc., Vol. 110, No. 13, 1988 4193

Table XI. Summary of Principal Component Analysis for the Parameter Sets[a]

	connectivity indices	Wiener indices	ad hoc descriptors	information indices
$f1$	8.745 (48.6)	5.187 (74.1)	4.918 (61.5)	4.550 (91)
$f2$	4.712 (74.8)	0.981 (88.1)	2.102 (87.8)	0.418 (99.4)
$f3$	1.556 (83.4)	0.642 (97.3)	0.534 (94.4)	0.027 (99.9)
$f4$	1.276 (90.5)	0.156 (99.5)	0.269 (97.8)	0.005 (99.99)
$f5$	0.672 (94.2)	0.032 (99.98)	0.108 (99.1)	0.0003 (100)
$f6$	0.448 (96.7)	0.001 (100)	0.059 (99.9)	
$f7$	0.279 (98.3)	0.00003 (100)	0.008 (99.98)	
$f8$	0.166 (99.2)		0.001 (100)	
$f9$	0.074 (99.6)			
$f10$	0.035 (99.8)			

[a] Eigenvalues (percent variance reproduced).

In contrast, $f2$ is entirely unrelated to V_B ($r^2 = 0.006$). Its equations are dominated by shape-dependent terms and higher order cluster terms (eq 8 and 9).

$$f2 = 2.8\ (\pm1.4) + 0.04\ (\pm0.01)T_3{}^2 - 1.4\ (\pm0.8)T_m +$$
$$0.2\ (\pm0.1)T_m{}^2 \quad (8)$$

$$n = 51 \qquad r^2 = 0.303 \qquad s = 0.9$$

$$f2 = 1.6\ (\pm0.6) + 3.4\ (\pm0.7)^6\chi_c - 0.6\ (\pm0.2)^2\chi +$$
$$2.0\ (\pm1.0)^4\chi_c \quad (9)$$

$$n = 51 \qquad r^2 = 0.414 \qquad s = 0.8$$

Factor $f3$ is also unrelated to V_B ($r^2 = 0.001$). Its equations show influences from both mass- and path-dependent terms (eq 10 and 11).

$$f3 = 0.47\ (\pm0.26) + 0.61\ (\pm0.05)p_3 - 0.056\ (\pm0.006)N_c{}^2 -$$
$$0.05\ (\pm0.01)T_m{}^2 \quad (10)$$

$$n = 51 \qquad r^2 = 0.784 \qquad s = 0.5$$

$$f3 = -0.5\ (\pm0.3) + 1.7\ (\pm0.1)^3\chi_p - 0.28\ (\pm0.04)1/\chi_t -$$
$$0.7\ (\pm0.1)^3\chi_c \quad (11)$$

$$n = 51 \qquad r^2 = 0.785 \qquad s = 0.5$$

It is also possible to apply PCA to the *parameter sets* used for these C_2–C_9 alkanes. The results are shown in abbreviated form in Table XI. Application of several criteria noted under Methods indicates that in terms of their major variances the Wiener indices span 3 dimensions, the connectivity indices, 6 or 7 dimensions, the information indices, 2 dimensions, and the ad hoc descriptors, 3 or 4 dimensions.

Discussion

The regression equations in Tables VI and VII provide a basis for the estimation of missing values for most of the physical properties examined here. For most properties except mp the accuracies of the equations should be sufficient for many practical purposes. The equations reflect the degree to which each property can be represented in an additive/constitutive manner based on the present descriptors.

The ad hoc descriptor equations (Table VII) display the relative influences of molecular mass (N_c), branching (T_m), and steric factors (T_3, p_3). Clearly, mass is overwhelmingly the dominant influence. Branching and steric differences cause smaller alterations in the average intermolecular distances and corresponding alterations of the intermolecular forces.[15] Branching, e.g., sequesters parts of the molecule and thereby prevents close contact with neighboring molecules. Since intermolecular interaction energies fall sharply as $1/r^6$, this decreases the cohesive forces experienced by such molecules. Steric crowding causes a small reduction in the molar volume (see the negative coefficient for p_3 in the MV expression). Therefore properties such as bp, HV, TC, and ST, which depend directly on the strength of intermolecular forces, tend to display positive dependences on N_c, negative dependences on T_m, and positive dependences on p_3 or T_3. As a rule—unless they are dominant in magnitude—the quadratic and inverse terms account for nonlinear dependences and may take

either sign. The critical pressure (PC) decreases as the intermolecular forces increase and shows a negative dependence on N_c (its much smaller dependences on p_3 and T_3 do not follow this pattern). The signs and relative magnitudes of these influences are more completely illustrated in Table VIII.

The failure of the present parameters to model the melting points of these hydrocarbons is not surprising and demonstrates the greater subtlety of the melting transition compared to, e.g., the boiling and the critical transitions.[15,22] The latter processes depend, in a more or less straightforward manner, on complete disruption of intermolecular forces and thus on the strengths of these forces. Conversely, the melting transition maintains a condensed phase and involves disruption of specific intermolecular fits and geometric arrangements that lie beyond the scope of the present parameters. Results of the factor analysis further emphasize this distinction and show melting to lie on a dimension almost entirely independent of the other properties. Modeling of this dimension (eq 8 and 9) shows it to depend on shape-dependent, as contrasted to mass-dependent, terms. In some cases a comparative approach has been used in treating melting points.[54] Recently, Hanson and Rouvray[55] have reviewed studies of melting points and achieved successful correlations with topological indices for normal alkanes. They indicate that their method can be extended also to the branched alkanes.[56] Kier[57,58] has recently described shape indices that hold promise for the description of melting points.[59]

All four structural parameter sets tested were reasonably successful in accounting for the physical properties of these alkanes, but the connectivity indices and ad hoc descriptors generally provided the most accurate models (Table IV). In part, this reflects the greater numbers of parameters in these two sets. Nonetheless, we conclude that these latter indices form well-chosen sets of structural descriptors with which to account for the physical properties of the normal and branched alkanes. When developing the form of the molecular connectivity index, Randić emphasized that this form was a solution to the bond-ordering problem for physical properties.[20,60] In a comparison of six topological indices, Razinger et al.[61] found the connectivity index to be the most effective in representing several properties of the alkanes. The success of connectivity indices in the present example and in an enormous variety of other applications[8-10,13-15,20,22,41,62,63] speaks strongly for the form chosen. This view is reinforced by the factor analysis, which reveals that the connectivity indices and ad hoc descriptors span more major, independent dimensions than do the other parameter sets. Although it must be emphasized that a parameter set with more major dimensions will not necessarily contain dimensions physically relevant to a set of properties,[64] a priori a set with more dimensions may be expected to have a better chance of containing these relevant dimensions than a set with fewer dimensions. As an initial condition, a parameter set should contain at least as many dimensions as do the properties to be modeled (here two or three) in order to have a good chance of success. In a relevant study Basak et al. have examined 90 topological indices using PCA on a diverse data set of 3692 structures and found 10 significant principal components.[50]

Recently, attention has turned to the use of alternative expressions, such as the Walker form[65]

(54) Charton, M.; Charton, B. I. In *QSAR in Design of Bioactive Compounds*; Prous: Barcelona, 1984; p 41.
(55) Hanson, M. P.; Rouvray, D. H., to be submitted for publication.
(56) Rouvray, D. H., presonal communication.
(57) Kier, L. B. *Quant. Struct.-Act. Relat. Pharmacol., Chem. Biol.* **1985**, *4*, 109; **1986**, *5*, 1, 7; **1987**, *6*, 8.
(58) Kier, L. B. *Acta Pharm. Jugosl.* **1986**, *36*, 171.
(59) Kier, L. B., personal communication.
(60) Trinajstic, N.; Randić, M.; Klein, D. J. *Acta Pharm. Jugosl.* **1986**, *36*, 267.
(61) Razinger, M.; Chretien, J. R.; Dubois, J. E. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 23.
(62) Jain, D. V. S.; Singh, S.; Gombar, V. *Proc.—Indian Acad. Sci., Chem. Sci.* **1984**, *93*, 927.
(63) Jain, D. V. S.; Singh, S.; Gombar, V. *Indian J. Chem., Sect. A* **1985**, *24A*, 545.
(64) Jolliffe, I. T. *Appl. Statist.* **1982**, *31*, 300.

4194 *J. Am. Chem. Soc., Vol. 110, No. 13, 1988*

Needham et al.

$$\text{property} = a[\text{index}]^b \qquad (12)$$

where a and b are coefficients to be determined, in structure–property analysis. Rouvray and Pandey[66] have employed the Wiener index and Walker expressions in an analysis of the boiling points of normal alkanes. They found different dependences for the lower and higher molecular weight alkanes, which they attributed to the effects of folding in the longer chains. Similar arguments have appeared in other contexts.[67] In the multiple linear regression analysis, these longer chain effects are represented to some extent in the nonlinear terms such as $1/N_c$ and N_c^2 and by p_3 and higher order path terms.

In the present study, limited to the physical properties of a class of nonpolar compounds, a single major factor was dominant for seven of the eight properties examined (Table IX). The most reasonab e interpretation of this factor is as a mass- or bulk-related dimension associated with intermolecular dispersion forces. This interpretation is consistent with the strong correlation of this factor with molecular volume[68] and with its expansion in terms of *ad hoc* descriptors and connectivity indices (eq 6 and 7). It is also reflected in the *f*1 factor *scores* (not shown[69]), which are most negative for the lower molecular weight, branched compounds and most positive for *n*-nonane and several lightly branched nonanes. Meyer and co-workers[70,71] have shown that dispersion forces dominate the liquid properties even of relatively polar compounds, and such forces are surely of overwhelming importance for the alkanes examined in the present investigation. A second factor, associated with the melting points, is correlated with features of molecular shape and steric fit (eq 8 and 9). This factor is only of minor importance for properties other than mp. A third factor, independent of molecular volume and difficult to characterize otherwise, appears to account for a smaller fraction of the property variances not represented in the first two factors.

In an important study of the physical properties of a diverse set of 114 compounds, Cramer[47] used PCA to identify two major dimensions, which he termed *B* and *C*, and three minor dimensions, termed *D*, *E*, and *F*, for the properties. The major dimensions *B* and *C* accounted for about 96% of the variance in six of the properties. Cramer associated dimension *B* with "some aspect of molecular *bulk*". The largest molecules in the set had the highest *B* scores, and the smallest, the lowest *B* scores. *C* was identified as a polar dimension related to molecular "cohesiveness;" because of the orthogonality of the PCA axes, this second axis was more exactly described as cohesiveness relative to the given molecular bulk.[47] Dimension *D* was associated with "dispersion interactions" and molecular "deformability", and dimensions *E* and *F*, with minor, ideosyncratic aspects of molecular behavior.

Cramer also applied these concepts to a set of 139 additional compounds and derived additive-constitutive models permitting calculation of the factor scores from molecular structure.[47]

On the basis of the present results and an examination of Cramer's results, we offer the following reinterpretation of Cramer's dimensions in terms of physical interactions. The bulk dimension *B* is clearly a measure of molecular *dispersion* interactions. As noted by Cramer, *B* is strongly associated with molecular mass/volume, a feature typical of dispersion interactions. Cramer's linear fragment model for calculation of molecular *B* values is entirely consistent with this; e.g., the values for atoms $H \approx F < Cl < Br < I$ are in the order of the atomic polarizabilities. Even the negative dependence on alkane branching noted in our work is reflected in Cramer's alkane *B* scores. The fundamental dependencies on *B* of the physical properties are also typical of dispersion interactions (cf. Table VI of ref 47 and our Table VIII). Thus, Cramer's *B* dimension corresponds to the major dimension found in the present study. We see no compelling reason to associate Cramer's minor *D* dimension with dispersion interactions, although this dimension may represent a minor adjustment to either the *B* or *C* dimensions. *C* is certainly a "polar cohesiveness" measure, but it is dominated by hydrogen-bonding contributions. The most positive *C* values in the original compound set occur for acetamide, water, acetic acid, methanol, and ethanol—all strong hydrogen-bonding agents; the most negative *C* values are for branched hydrocarbons, such as 2,2-dimethylbutane. Of the additional 139 compounds examined by Cramer, glycerol, hydrazine, and ammonia have the most positive *C* values, and hexadecane has the lowest. This influential hydrogen-bonding component is also reflected in Cramer's linear fragment model[47] for *C* where the largest positive contributions are from –OH, –CN, >C=O, and –NH$_2$ groups. This polar, hydrogen-bonding dimension is necessarily absent from our present analysis of the alkanes.

In conclusion, we have applied graph-theoretical molecular modeling techniques to eight representative physical properties of the linear and branched alkanes through the nonanes. With the notable exception of the melting points, connectivity indices and ad hoc descriptors are found to be very useful parameter sets for modeling these properties. Factor analysis reveals that most of these properties are dominated by a single mass/volume-related dimension, reasonably associated with intermolecular dispersion interactions. Melting points are associated with a shape-dependent dimension not well modeled by the indices examined.

(65) Walker, J. *J. Chem. Soc.* **1894**, *65*, 725.
(66) Rouvray, D. H.; Pandey, R. B. *J. Chem. Phys.* **1986**, *85*, 2286.
(67) Ungar, G.; Stejny, J.; Keller, A.; Bidd, I.; Whiting, M. C. *Science (Washington, D.C.)* **1985**, *229*, 386.
(68) Charton, M. *Top. Curr. Chem.* **1983**, *114*, 107.
(69) Available from P.G.S.
(70) Meyer, E. F.; Wagner, R. E. *J. Phys. Chem.* **1966**, *70*, 3162.
(71) Meyer, E. F.; Renner, T. A. *J. Phys. Chem.* **1971**, *75*, 642.